DOCUMENT RESUME

ED 448 188 TM 032 149

AUTHOR McLean, James E.; O'Neal, Marcia R.; Barnette, J. Jackson
TITLE Are All Effect Sizes Created Equal?
PUB DATE 2000-11-00
NOTE 16p.; Paper presented at the Annual Meeting of the Mid-South Educational Research Association (28th, Bowling Green, KY, November 15-17, 2000).
PUB TYPE Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Effect Size; National Surveys; Norm Referenced Tests; *Scores
IDENTIFIERS *Normal Curve Equivalent Scores

ABSTRACT

This study compared effect sizes applied to raw, scaled, and normal curve equivalent (NCE) data. Recommendations for the interpretation of effect sizes vary. For example, some authors suggest that an effect size below 0.50 is small, between 0.50 and 1.00 is moderate, and above 1.00 is large. These are products of the criterion formally used by the U.S. Department of Education's Joint Dissemination Review Panel and the Program Effectiveness Panel. it is clear from the context of these articles that it is assumed that they were dealing with raw scores or scaled scores, not NCEs. NCE scores for individual students and, particularly, mean NCE scores for schools would not be expected to change from year to year without some type of intervention. This study computed effect sizes for the raw, scaled scores, and NCEs by school for grades 4, 6, and 8 on a national norm-referenced test for 749, 574, and 464 schools respectively representing 120,149 students. The results show that, as expected, the effect sizes for NCE scores were lower than those for raw and scaled scores. These results suggest that when rules-of-thumb for effect sizes are presented, they should take into account the type of metric on which it is being applied. (Contains 3 figures, 4 tables, and 23 references.) (SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 448 188

# Are All Effect Sizes Created Equal?

James E. McLean
East Tennessee State University

Marcia R. O'Neal
University of Alabama at Birmingham

J. Jackson Barnette
University of Iowa

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

J. McLean



TM032149

Paper Presented
At the Annual Meeting
Of the
Mid-South Educational Research Association
Bowling Green, KY
November 2000

BEST COPY AVAILABLE

Abstract

The publication of the Glass et al on meta-analysis created a cottage industry in effect size computation. The recent debate over statistical significance testing has reinforced the interest in effect size. Much of the current knowledge about effect sizes comes from the work of Cohen presented in his text on power analysis. However, the literature makes no distinctions among effect sizes based on the data metric upon which they are applied. The purpose of this study was to compare effect sizes applied to raw, scaled, and normal curve equivalent (NCE) data.

Recommendations for the interpretation of effect sizes vary. For example, some authors suggest that an effect size below .50 is small, between .50 and 1.00 is moderate, and above 1.00 is large. These are products of the criterion formally used by the U.S. Department of Education's Joint Dissemination Review Panel (JDRP) and the Program Effectiveness Panel (PEP). It is clear from the context of these articles that it is assumed that they were dealing with raw scores or scaled scores, not NCEs. NCE scores for individual students and, particularly, mean NCE scores for schools would not be expected to change from year to year without some type of intervention.

This study computed gain effect sizes for the raw, scaled scores, and NCEs by school for grades 4, 6, and 8 on a national norm-referenced test for 749, 574, and 464 schools respectively representing 120,149 students. The effect sizes were compared for each type of score. The results showed that, as expected, the effect sizes for NCE scores were lower than those for raw and scaled scores. These results suggest that when rules-of-thumb for effect sizes are presented, they should take into account the type of metric upon which it is being applied.

## Are All Effect Sizes Created Equal?

Many researchers recommend using effect sizes to ascribe the practical meaning of results (e.g., Cohen, 1988; Glass, McGaw, & Smith, 1981; McLean, 1983; McLean & Ernest, 1998; Slavin & Fashola, 1998). Recommendations vary in terms of the interpretation of effect sizes. For example, McLean (1995, p. 40) suggests that an effect size below .50 is small, between .50 and 1.00 is moderate, and above 1.00 is large. McLean based his criterion on that formally used by the U.S. Department of Education's Joint Dissemination Review Panel (JDRP) and the Program Effectiveness Panel (PEP). Slavin and Fashola (1998) suggested that an effect size equal to or greater than .25 should be considered evidence of effectiveness. All of these approaches were based on experience and judgment. An exception to this approach was Barnette and McLean (1999, November). They demonstrated empirically that it takes an effect size of at least .50 to be reasonably sure that the difference did not happen by chance. They suggested (Barnette and McLean, 2000, April) using a statistical significance test before applying an effect size formula to protect from interpreting random effect sizes. However, even Barnette & McLean (1999, November; 2000, April) did not take into account the metric used in reporting the results in their analyses. The purpose of this study was to extend this research by comparing effect sizes using raw, scaled, and normal curve equivalent (NCE) data from a nationally normed standardized test given statewide.

### Background

While the term, *effect size*, is rather recent, the concept has been around for many years. There is evidence that pioneer statisticians in the first half of the 20th century recognized the need to consider the meaningfulness of the results beyond just their statistical significance (e.g., Fisher, 1938). The first documented formal uses of effect size estimates in education came as a consequence of the Elementary and Secondary Education Act of 1965. This Act provided for the dissemination of innovative educational programs that were certified to be effective. To provide this certification, the Joint Dissemination Review Panel (JDRP) was established consisting of personnel from the then, Office of Education, and National Institute of Education. Before a program

could be considered for dissemination funding by the National Diffusion Network (NDN), it had to be approved by the JDRP.

While the JDRP had numerous criteria that could be categorized under replicability and effectiveness, a key component of the effectiveness criteria was effect size. It became clear that it was very difficult for a project to be approved by the JDRP if it could not demonstrate an effect size of at least 1.00. The 1.00 effect size became the defacto criterion for an effective project.

A number of publications have addressed this issue since that time. Probably Cohen's (1988) and Glass' (e.g., Glass, 1976, 1978; Glass, et al., 1981) works have been the most influential. Cohen popularized the use of effect size reporting in almost all statistical analyses. The Glass, et. al publications on meta-analysis expanded the use of effect size from a recommended outcome to report with a study to the dependent variable in a new study. In 1983 (March), McLean recommended extending this approach to determining the effectiveness of NDN programs by using effect sizes to estimate the overall effectiveness of the adoptions of NDN programs.

Barnette and McLean (1999, November) conducted a Monte Carlo study using effect size as the outcome variable where they generated thousands of sets of data under a "no difference" condition in the populations. They estimated the average effect size when there was actually no difference in the population. Thus, they obtained effect sizes one might obtain by chance. A conclusion of that study is that chance effect sizes of .50 are common.

In the cases cited above, no assumptions were made about the type of scores used to compute the effect sizes. They could have been based on raw scores, scaled scores, or any other type of metric. None of the publications suggested that they were considering normal curve equivalent scores (NCEs). NCE scores are intervalized percentile rank scores and, as such, are not expected to change for year to year for individual students and, even more so, for mean NCE scores for schools without some type of intervention. An NCE score is a normative measure comparing relative performance to a norming population. If a student (or school) maintains his/her (or its) place in the norming

population, the NCE score would remain constant. Thus, any year-to-year increase in NCE is important.

Perhaps no one has had a greater impact on the use of effect sizes than Cohen (1969, 1988) through his work on power analysis. In these publications, Cohen suggested general guidelines for levels of effect size. These are .2 for small effect, .5 for medium effect, and .8 for large effect.

A broader debate on the use of statistical significance testing emerged from Cohen's power analysis works. Kaufman (1998) indicated that the "controversy about the use or misuse of statistical significance testing has been evident in the literature for the past 10 years and has become the major methodological issue of our generation" (p. 1). The debate has ranged from those who recommend the elimination of statistical significance testing (e.g., Carver, 1978, 1993; Nix & Barnette, 1998) to those who staunchly support it (e.g., Frick, 1996; Levin, 1993, 1998; McLean & Ernest, 1998). However, even those who defend statistical significance testing indicate that significant results should be accompanied by a measure of practical significance. The leading method of reporting practical significance is through the provision of an effect size estimate (Kirk, 1996; McLean & Ernest, 1998; Robinson & Levin, 1997; Thompson, 1993). Unfortunately, the meaning of effect size is still open to question.

## Method

This study used the data from a state-wide testing program. The test was the Ninth Edition of the Stanford Achievement Test. Results from the Spring 1998 administration and the Spring 1999 administration were used in the analysis. From the spring 1998 files, all students in Grades 4, 6, and 8 (except those with blank student numbers or missing Total Reading NCEs) were selected and their school, grade, student number, and Reading Total raw scores, scaled scores, and NCEs were obtained (n = 150,071). From the spring 1999 files, all students in Grades 5, 7, and 9 (except those with blank student numbers or missing Total Reading NCEs) were selected their school, grade, student number, and Reading Total raw scores, scaled scores, and NCEs were obtained (n = 153,115).

Some general definitions of the three types of scores might be in order. Raw scores, as everyone knows, are found by adding up the number of correct answers on a test. Scaled scores, in this case, are found by doing a linear rescaling of the raw scores across all of the grades covered by the test. The scaled scores range from appoximately 100 to 900. Normal curve equivalent (NCE) scores need a little more explanation. They were developed by RMC Research Corporation in 1976 to measure the effectiveness of the Title I Program across the United States. Essentially, NCE scores are intervalized percentile ranks to render them appropriate for parametric statistical analyses. The 1st, 50th, and 99th percentile rank scores and NCE scores are the same, but the scores in between these are different. This is because the NCE scores have been spaced at equal intervals. NCE scores have a mean of 50 and a standard deviation of approximately 21.06 (Wothen, White, Fan, & Sudweeks; 1999).

The two sets of data were matched on student number, eliminating any with other erroneous student numbers and/or missing data resulting in n = 120,149 cases with complete pretest (spring 1998) and posttest (spring 1999) data. Using these matched cases, pre and post sample sizes, means, and standard deviations for each type of score (raw, scaled, and NCE) were computed for each school and grade combination. Then if the sample size for a school/grade combination was 10 or greater, an effect size was calulated for each type of score (raw, scaled, and NCE) using the following formula:

$$\frac{(Posttest\ \ Mean) - (Pretest\ \ Mean)}{(Pretest\ \ SD)}$$

Finally, the differences between pairs of effect sizes, frequencies on effect sizes and the differences, correlations between pairs of effect sizes, and scatter plots between pairs of effect sizes were computed. There is a total of 1,787 sets of ns, means, standard deviations, effect sizes, and effect size differences – representing 1,787 school and grade combinations and 120,149 students.

## Results

The results are presented based on the differences between the effect sizes for each type of scale and the relationships among the effect sizes for each type of scale. Both tables and figures are used. Table 1 provides information on the grades represented, number of schools included, and the number of students in each of these grades.

Table 1

Sample Size Information

| Pretest Grade | Number of Schools | Number of Students | Number of Students for School/Grade Combinations |
|---|---|---|---|
| 4 | 749 | 41,178 | 10 - 307 |
| 6 | 574 | 40,589 | 10 - 447 |
| 8 | 464 | 38,382 | 10 - 428 |
| | 1,787 | 120,149 | 10 - 447 |

Thus, the study represents raw, scaled, and NCE effect size scores computed for 1,787 school and grade combinations based on 120,149 students in pretest Grades 4, 6, and 8.

Table 2 presents the descriptive statistics for the various effect sizes and differences between the effect sizes. The effect sizes in order of size are the scaled score, raw score, and NCE. Thus, the scaled and NCE score effect sizes represent the greatest difference.

Table 2

Descriptive Statistics for Effect Sizes and Effect Size Differences for Raw, Scaled, and NCE Scales for 1,787 Schools

| Variable | Mean | SD | Minimum | Maximum |
|---|---|---|---|---|
| Raw Score | -0.05 | 0.2964 | -1.85 | 2.05 |
| Scaled Score | 0.29 | 0.3135 | -1.24 | 2.91 |
| NCE Score | -0.17 | 0.2596 | -1.61 | 2.45 |
| Raw - Scaled | -0.34 | 0.1533 | -1.07 | -0.94 |
| Raw - NCE | 0.12 | 0.1843 | -0.40 | 0.94 |
| Scaled - NCE | 0.46 | 0.1326 | 0.22 | 1.23 |

Table 3 extends this information by presenting the effect size results in a frequency distribution categorized by the size of the effect size.

Table 3

Raw, Scaled, and NCE Effect Sizes Categorized by Size

| Category of Strength | | Type of Score | | | | | |
|---|---|---|---|---|---|---|---|
| | | Raw Score | | Scaled Score | | NCE | |
| Label | Values | Frequency | Percentage | Frequency | Percentage | Frequency | Percentage |
| 0 or Decline | 0 or less | 1,056 | 59.09% | 382 | 21.38% | 1,404 | 78.57% |
| Very Small | .01 - .24 | 426 | 23.84% | 323 | 18.08% | 324 | 18.13% |
| Small | .25 - .49 | 260 | 14.55% | 662 | 37.05% | 42 | 2.35% |
| Moderate | .50 - 1.00 | 42 | 2.35% | 392 | 21.94% | 12 | 0.67% |
| Large | > 1.00 | 3 | 0.17% | 28 | 1.57% | 5 | 0.28% |
| Total | | 1,787 | 100.00% | 1,787 | 100.00% | 1,787 | 100.00% |

Table 4 presents the correlations of effect sizes among the score types.

Table 4

Correlations of Effect Sizes Among Raw, Scaled, and NCE Score Types

| | Raw Score | Scaled Score | NCE Score |
|---|---|---|---|
| Raw Score | 1.00 | .88* | .79* |
| Scaled Score | | 1.00 | .91* |
| NCE Score | | | 1.00 |

* $p < .0001$

While the greatest effect size difference is between scaled and NCE effect size scores, they also correlate the highest ( $r = .91$) accounting for 83% of shared variance. Raw and NCE scores correlate the least at $r = .79$ representing 62% of shared variance.

The scatter plots for the three correlations are presented in Figures 1, 2, and 3.

Figure 1. Scatter plot of Raw vs. Scaled Scores. Note that 528 observations are hidden. The scales is A = 1 score, B = 2 scores, etc.

Figure 2. Scatter plot of Raw vs. NCE Scores. Note that 660 observations are hidden. The scales is A = 1 score, B = 2 scores, etc.

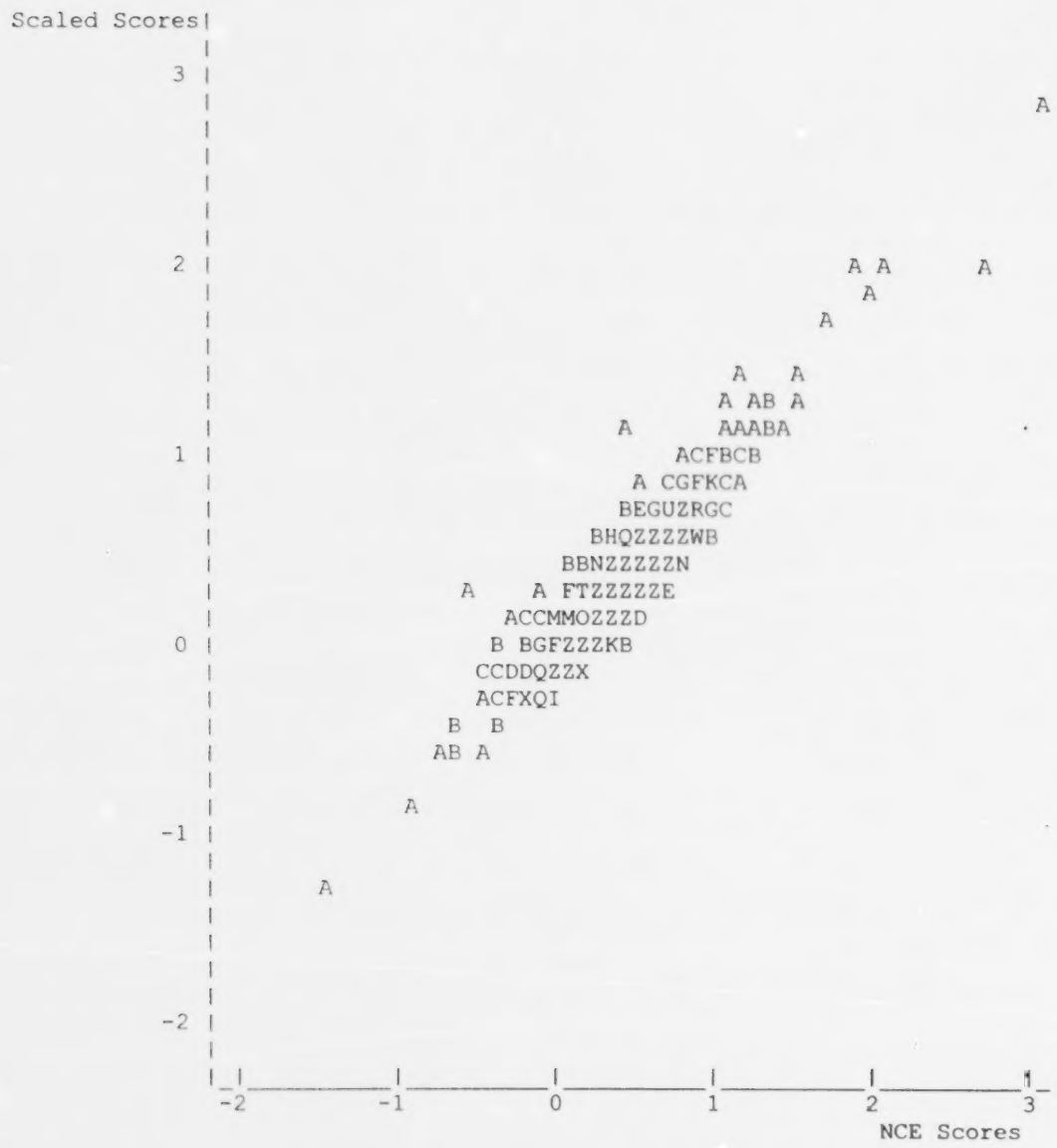


Figure 3. Scatter plot of Scaled vs. NCE Scores. Note that733 observations are hidden. The scales is A = 1 score, B = 2 scores, etc.

## Discussion and Conclusions

At least for this sample, the greatest difference was observed between the scaled and NCE effect size scores. Since NCEs are computed based on performance relative to each year's population, students, on the average, would be expected to remain stable from year to year if no special intervention was in place, resulting in an effect size of 0.0. However, in this case, the actual NCE effect size was -.017, suggesting that the group, on the average, did not progress at the level of its norming population. On the other hand, the scaled score effect size was 0.29 suggesting the students did make some improvement. Scaled scores are designed to index growth from year to year. At first blush, this seems to be in conflict with the negative raw score effect size (-0.05). However, it should be noted that since this was based on different tests given in 1998 and 1999, the number of raw score items on the pre and post tests may have differed.

Less information can be concluded from the correlations. For example, the fact that the highest correlation was between the scaled and NCE effect sizes has few implications for the purposes of this study. It probably suggests that higher achieving students demonstrated both greater gains based on scaled scores as well as more improvement when compared to their norming population than did lower achieving students.

The primary conclusion of this study is that the metric or type of score does make a difference when computing effect sizes. Thus, we should not have a one-size-fits-all rule-of-thumb to interpret effect sizes. We should take the type of score along with other factors into account when we interpret an effect size.

References

Barnette, J. J., & McLean, J. E. (1999, November). *Empirically based criteria for determining meaningful effect size.* Paper presented at the annual meeting of the Mid-South Educational Research Association, Point Clear, AL.

Barnette, J. J., & McLean, J. E. (2000, April). *Use of significance test as protection against spuriously high standardized effect sizes: Introduction of the protected effect size.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Carver, R. P. (1978). The case against statistical significance testing. *Harvard Educational Review, 48*, 378-399.

Carver, R. P. (1993). The case against statistical significance testing, revisited. Journal of Experimental Education, 61(4), 287-292.

Cohen, J. (1969). *Statistical power analysis for the behavioral sciences.* New York: Academic Press.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Fisher, R. A. (1938). *Statistical methods for research workers* (7th ed.). Edinburgh, Scotland: Oliver and Boyd.

Frick, R. W. (1996). The appropriate use of null hypothesis testing. *Psychological Methods, 1*(4), 379-390.

Glass, G. V., McGaw, B., & Smith, M. L. (1981). *Meta-analysis in social research.* Beverly Hills, CA: Sage Publications.

Glass, G. V (1978). Integrating findings: The meta-analysis of research. *Review of Research in Education, 5*, 351-379.

Glass, G. V (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher, 5*, 3-8.

Kaufman, A. S. (1998). Introduction to the special issue on statistical significance testing. *Research in the Schools, 5*(2), 1.

Kirk, R. E. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement, 56*(5), 746-759.

Levin, J. R. (1993). Statistical significance testing from three perspectives. *Journal of Experimental Education, 61*(4), 378-381.

Levin, J. R. (1998). What if there were no more bickering about statistical significance tests? *Research in the Schools*, *5*(2), 43-53.

McLean, J. E. (1983). *A meta-analysis approach to impact evaluation of adoptions*. Paper presented at the National Diffusion Network Regional Meeting, Memphis, TN. (ERIC Document Reproduction Service No. ED 242 744).

McLean, J. E. (1995). *Improving education through action research: A guide for administrators and teachers*. Thousand Oaks, CA: Corwin Press, Inc.

McLean, J. E., & Ernest, J. M. (1998). The role of statistical significance testing in educational research. *Research in the Schools, 5*(2), 15-22.

Nix, T. W., & Barnette, J. J. (1998). The data analysis dilemma: Ban or abandon. A review of null hypothesis significance testing. *Research in the Schools, 5*(2), 3-14.

Robinson, D. L., & Levin, J. R. (1997). Reflections on statistical and substantive significance, with a slice of replication. *Educational Researcher, 26*(5), 21-26.

Slavin, , & Fashola, (1998). *Show me the evidence: Proven and promising programs for America's schools*. Thousand Oaks, CA: Corwin Press.

Thompson, B. (Guest Ed.). (1993). Statistical significance testing in contemporary practice [Special Issue]. *The Journal of Experimental Education, 61*(4).

Worthen, B. R., White, K. R., Fan, X., & Sudweeks, R. R. (1999). Measurement and assessment in schools (2nd ed). New York: Longman.

# END

U.S. Dept. of Education

Office of Educational
Research and Improvement (OERI)

# ERIC

Date Filmed
May 25, 2001